Symbiotic AI for Cyber Social Security: A Human-Centered SaaS Against Cyberbullying

Andrea Chezzi[1], Simone Lupo[1], Alessia Anna Catalano[1], Roberto Vadacca[1], Christian Catalano[2], Giorgio Fontana[1] and Luca Mainetti[1,†]

[1] *University of Salento, Dept. of Innovation Engineering, Via per Monteroni 165, Lecce, 73100, Apulia, Italy*
[2] *University of Bari "Aldo Moro", Dept. of Engineering, Bari, 70121, Apulia, Italy*

Abstract

The rapid expansion of digital communication platforms has intensified the emergence of cyber-social harms, including cyberbullying, hate speech, online gender-based violence, and manipulative behaviors. These phenomena exploit the speed, anonymity, and scalability of networked interactions, with severe psychological and social consequences for affected individuals and communities. Traditional moderation approaches remain insufficient due to the multimodal nature of harmful content and the continuous evolution of abuse patterns.

This paper introduces a human-centered Software-as-a-Service (SaaS) platform designed to support Cyber Social Security (CSS) by detecting and mitigating cyber-social threats across diverse media formats. Leveraging Symbiotic AI principles, the system integrates multimodal analysis — combining natural language processing, visual recognition, audio and meme interpretation, and contextual metadata — to identify subtle and emerging forms of online aggression. Human-in-the-loop feedback, explainable decision-making, and policy compliance components ensure trustworthy, ethically aligned deployment at scale.

Results from real-world platforms such as Reddit and Telegram demonstrate strong multimodal classification performance while highlighting current limitations in datasets, cultural nuances, and generalization. This work contributes a practical and extensible framework for safer digital ecosystems, supporting timely intervention and strengthening collective resilience against cyber-social harms.

Keywords

Cyber Social Security, Multimodal Content Moderation, Human-Centered AI, Online Safety, Symbiotic AI, Explainable AI, Cyberbullying Detection

CEUR-WS.org/Vol-4329/paper9.pdf

1. Introduction

Digital communication platforms have become integral to everyday life, enabling global connectivity, information exchange, and new forms of social interaction. However, these same infrastructures have amplified harmful behaviors — from cyberbullying to coordinated harassment — giving rise to a growing spectrum of cyber-social threats. The consequences extend beyond the screen, affecting victims' mental health, social belonging, and personal safety.

Cyberbullying, in particular, has proven challenging to define and regulate due to its constantly shifting manifestations. While rooted in traditional bullying — repeated, intentional aggression targeting individuals with limited ability to defend themselves — online dynamics introduce novel elements: anonymity, virality, persistence, and multimodal content forms. As such, a universal definition has yet to be reached, complicating consistent detection and reporting.

The increasing quantity and complexity of online content exceeds the capacity of human moderators and platform governance mechanisms. Harmful behavior may remain subtle or coded, hidden among

1st Workshop on Symbiotic AI for Cyber Social Security: Human–AI Collaboration for Social Resilience and Trustworthy Digital Ecosystems (SAI4CCS 2026), co-located with the 2nd Joint National Conference on Cybersecurity (ITASEC & SERICS 2026), February 09–13, 2026, Cagliari, Italy

*Corresponding author.
†These authors contributed equally.
andrea.chezzi@unisalento.it (A. Chezzi); simone.lupo@unisalento.it (S. Lupo); alessiaanna.catalano@unisalento.it (A. A. Catalano); roberto.vadacca@unisalento.it (R. Vadacca); christian.catalano@uniba.it (C. Catalano); giorgio.fontana@unisalento.it (G. Fontana); luca.mainetti@unisalento.it (L. Mainetti)

© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2026-11-14

memes, irony, emojis, or synchronized actions across platforms. Automated, scalable, and context-aware solutions are urgently needed to support early detection and protection measures.

To address these challenges, this work proposes a multimodal SAI platform that assists moderators, institutions, and digital providers in recognizing and mitigating abusive content in real time. The platform processes text, images, videos, audio streams, and interaction metadata, enabling comprehensive detection aligned with societal and regulatory priorities.

SAI represents a specific and advanced approach within human-centred artificial intelligence (HCAI) that goes beyond traditional automation and standalone decision-making systems. In the context of HCAI, SAI systems are designed to foster an ongoing, collaborative relationship between human intelligence and artificial intelligence such that both parties contribute to and benefit from the interaction. Rather than simply augmenting task performance, these systems aim at mutual enhancement: supporting human decision-making, preserving autonomy, and adapting continuously to human needs and cognitive models.

In the foundational framework proposed by Calvano and colleagues, SAI is conceptualized as an AI paradigm in which humans remain at the core of the design and development process. This involves principles such as transparency, fairness, appropriate levels of automation, and protection, all oriented toward sustaining trust and cooperation between users and AI. As a subtype of HCAI, SAI emphasizes ethical, interpretable, and context-aware interactions that respect human values and regulatory priorities, such as those articulated in the European Union's AI Act [1].

Importantly, SAI is not simply about making systems more usable; it requires deliberate design choices that keep humans involved in oversight and control, facilitate shared agency, and enhance users' abilities without replacing them. This symbiotic relationship allows AI to respond not only to explicit commands but also to evolving patterns of human behavior and context — a key strength when addressing nuanced and dynamic challenges such as online abuse detection.

By grounding the proposed multimodal platform in the principles of SAI, this work aims to ensure that technological scalability is matched with human-centred governance, ethical alignment, and real-time adaptability, ultimately enabling more resilient and socially aligned moderation strategies.

2. Motivation and Problem Statement

Online abuse today extends beyond explicit slurs or visible aggression. Modern cyberbullying and online hate often operate across multiple media channels: memes, manipulated images, voice messages, and livestream interactions. As shown by recent surveys, single-modality systems (e.g. text-only) are insufficient to capture the breadth of harmful content; multimodal approaches are required [2, 3].

Harmful communication is often ambiguous or coded through sarcasm, inside jokes, and dog-whistles, making intentions difficult to interpret. This is further complicated by language variation, cultural references, and shifting norms of expression — contextual sensitivity that many detection systems lack. [4, 5]

Public visibility amplifies power imbalances, enabling targeted harassment or public shaming that can spread broadly, persist over time, and inflict long-term social and psychological harm. This social dimension — including the dynamics of who is targeted and why — is often overlooked in purely algorithmic moderation systems [6].

Despite progress, existing moderation tools and research frameworks struggle for various reasons. Multimodal content such as images, video, and audio is still under-moderated due to limited analytical capabilities [7]. Contextual understanding of relationships, cultural nuance, and intent remains shallow, leading to misclassifications or missed threats [8, 9]. Manual review processes are slow, reactive, and resource-intensive, while linguistic and cultural variation increases the risk of excluding or misinterpreting minority communities [10]. These core challenges and their practical consequences are summarized in Table 1.

Table 1
Key challenges faced by current online abuse detection and moderation systems and their practical impacts.

Challenge	Impact
Limited multimodal analysis	Media such as images, videos, and audio remain under-moderated
Insufficient contextual reasoning	Intent, relationships, and cultural nuance are overlooked
High dependency on manual review	Moderation remains slow, reactive, and resource-intensive
Linguistic and cultural bias	Minority communities face misinterpretation or exclusion

3. Related Work on Multimodal Detection of Cyber-Social Harm

Early approaches to cyberbullying and hate speech detection predominantly focused on single modalities, especially text. Classical NLP systems relying on keyword spotting, bag-of-words representations or TF–IDF features combined with traditional classifiers were effective in identifying explicit insults and slurs, but performed poorly in cases where abusive intent is conveyed through irony, visual cues or multimodal interplay [11]. As online harms increasingly take the form of memes and visually augmented content, researchers have emphasized that abusive meaning often emerges only when image and text are interpreted jointly [12, 13]. The introduction of multimodal benchmark datasets—including the influential Hateful Memes dataset—showed that unimodal models frequently misclassify examples where one modality appears benign but contributes to harmful meaning when combined with another [11].

3.1. Evolution of Multimodal Fusion

Initial multimodal systems adopted *early fusion*, concatenating text and image features prior to classification. Although this approach enabled simultaneous observation of all modalities, its performance was limited by weak semantic alignment across high-dimensional feature spaces [14]. A shift toward *late fusion* improved modularity and made systems easier to train, yet these architectures still underperformed on content requiring fine-grained cross-modal reasoning.

Transformer-based multimodal architectures marked a considerable step forward. Models such as ViLBERT, UNITER and LXMERT employ intermediate fusion through cross-attention mechanisms to model relationships between visual regions and textual phrases, enabling the detection of cases where the text alone is misleading or benign [15]. Their superiority has been demonstrated through performance gains on datasets with visually dependent hate semantics, particularly memes designed to fool unimodal baselines [11, 16].

3.2. Multimodal Architectures for Cyberbullying and Hate Speech

The first practical implementations for cyberbullying detection typically used dual-branch architectures: CNN backbones such as VGG or ResNet encoded the visual modality, while Bi-LSTMs or GRUs processed textual content. Feature embeddings were concatenated to train a classifier. These systems consistently improved over text-only baselines, especially in tasks involving appearance-based shaming and meme-driven harassment [12]. As multimodal technologies matured, pre-trained contrastive vision–language encoders such as CLIP were integrated to enhance interpretability and robustness [15].

More recently, metadata and social interaction patterns have been added as complementary modalities. Graph Neural Networks capture relational signals such as repeated bullying by influential users or coordinated harassment campaigns [17]. When fused with multimodal text–image models, social-graph features significantly improve macro-F1 performance on real-world cyber-social harm datasets. Additionally, recent work has pushed beyond English and high-resource languages, demonstrating that multimodal fusion improves detection accuracy even in low-resource cyberbullying contexts [18].

Despite this progress, current systems remain challenged by evolving multimodal manipulations, cultural nuances, and cross-platform generalization. Ongoing research increasingly explores scalable, explainable and socially aware multimodal AI frameworks capable of collaborating with human expertise for reliable cyber-social harm mitigation.

4. Traditional Bullying vs. Cyberbullying: A Shift in Dynamics

Although cyberbullying stems from the same fundamental power imbalance as traditional bullying, it alters the dynamics through the affordances of digital environments. Traditional bullying typically requires physical co-presence and is constrained to bounded social settings such as schools or communities. Cyberbullying, by contrast, operates through ubiquitous digital channels, producing a sense of constant exposure and inescapability [19, 20]. Offline anonymity is generally limited, yet online pseudonymity enables identity manipulation, increasing impunity and the persistence of aggressive behavior [21].

Audience scale also differs markedly: traditional bullying is usually restricted to a local peer group, while cyberbullying can spread rapidly through networked publics, amplifying humiliation and reputational damage [22]. Furthermore, while the impact of offline events is often temporally bounded, digital content can be stored, replicated, and re-circulated indefinitely, creating long-term psychological and social consequences for victims [23]. The role of bystanders shifts as well; whereas offline peers may intervene directly, online bystanders may unintentionally reinforce abuse through passive engagement such as likes or shares, shifting responsibility toward platform governance and algorithmic content curation [24]. These structural differences make detection and intervention more complex, immediate, and consequential.

4.1. Psychological, Social, and Regulatory Implications

Cyberbullying produces severe and well-documented psychological and social consequences. Victims frequently experience chronic anxiety, hyper-vigilance, depressive symptoms, social withdrawal, diminished self-esteem, and, in acute cases, suicidal ideation [25, 26]. Unlike traditional forms of bullying, digital communication environments enable harassment to persist beyond physical or temporal boundaries. The continual accessibility of online platforms allows abusive behavior to occur at any time, creating a sustained sense of exposure and vulnerability. This pervasive and enduring nature of online aggression can generate levels of distress comparable to or exceeding those associated with offline bullying, effectively constituting a form of ongoing digital trauma that calls for proactive technological and institutional intervention [27].

These individual harms also translate into broader societal and governance challenges. As the scale and impact of cyberbullying become increasingly evident, governments and institutions have begun implementing regulatory and policy measures aimed at protecting users and promoting safer digital ecosystems. Legislative frameworks such as the Italian Law 71/2017, the General Data Protection Regulation (GDPR) provisions on sensitive personal data [28], and European initiatives including the Safer Internet Programme and the Better Internet for Kids strategy [29] emphasize rapid content removal, privacy-preserving procedures, and responsible platform accountability. In the United States, state-level regulations and school-based mandates encourage timely reporting mechanisms and collaborative human–AI assessment models [20]. In parallel, major social media platforms have adopted global governance standards centered on risk-based, transparent, and explainable moderation practices.

Despite these advances, significant challenges persist. Regulatory approaches remain fragmented across jurisdictions, enforcement and takedown processes are often slow or inconsistent, and automated detection systems continue to struggle with contextual subtleties such as irony, cultural specificity, and multimodal meme-based communication [30]. Furthermore, excessively restrictive moderation policies risk infringing upon freedom of judgement, underscoring the need for balanced solutions. These limitations highlight the importance of human-centred and SAI approaches that support, rather

than replace, human judgment, combining computational scalability with contextual sensitivity and ethical oversight.

5. Social Media as Cyber-Social Sensors

Digital platforms can be conceptualized as social sensors: socio-technical infrastructures that continuously generate signals about human behavior, collective emotions, and emerging tensions. Unlike physical sensors that measure environmental variables such as temperature or movement, social sensors capture user-generated content and interactions including posts, comments, images, voice messages, reactions, and group dynamics. This view aligns with prior work in computational social science and social sensing, which treats online platforms as large-scale observatories of societal dynamics and collective behavior [31, 32]. Within a Cyber Social Security (CSS) framework, these platforms operate as large-scale antennas capable of detecting early indicators of cyberbullying, hate speech, coordinated harassment, misinformation, and other cyber-social harms [6]. Appropriately designed AI systems can transform this noisy and heterogeneous data stream into actionable insights for prevention, mitigation, and evidence-based policy support.

5.1. Data Types Collected by Social Sensors

The proposed SaaS platform adopts an explicitly multimodal approach, exploiting the diverse types of data emitted by online platforms. Each modality demands specialized computational techniques to effectively identify signals associated with cyberbullying and related phenomena [33].

5.1.1. Textual Data

Text remains the dominant interaction format across most platforms, including posts, tweets, comments, replies, private messages, and forum discussions. Through Natural Language Processing (NLP), combined with sentiment and toxicity analysis, the system can detect insults, slurs, threats, and humiliation, as well as coordinated attacks such as dogpiling. Analysis of conversational structure further enables identification of recurring patterns targeting particular individuals or groups. Additionally, monitoring emerging hashtags, recurrent expressions, and linguistic cues supports early-warning mechanisms capable of flagging the onset of targeted campaigns [34, 35].

5.1.2. Visual Data: Images and Video

Platforms like Instagram, TikTok, Snapchat, and Facebook rely heavily on visual content, which has become a primary medium for cyberbullying and online hate. Relevant abusive behaviors include non-consensual sharing of photos, appearance-based mockery and body shaming, meme-based ridicule, and the disclosure of private imagery intended to shame individuals. Computer Vision methods enable recognition of harmful visual content by detecting explicit or suggestive imagery used for harassment, reading embedded abusive text via Optical Character Recognition (OCR), and characterizing malicious memes or reaction videos directed at vulnerable groups [7]. This capability is essential given that much contemporary aggression is encoded visually rather than verbally.

5.1.3. Audio and Voice Messages

Voice-based communication is increasingly prevalent in platforms such as Discord, Telegram, WhatsApp, Clubhouse, and online gaming environments. Through Automatic Speech Recognition (ASR) and audio analysis, the system can transcribe voice notes, detect verbal aggression including threats and humiliation, and evaluate prosodic features such as tone and intensity as additional indicators of hostility. Prior studies on speech-based toxicity and paralinguistic analysis demonstrate that acoustic cues can complement textual transcriptions for identifying harmful intent [36, 37].

5.1.4. Geolocated Data

While cyberbullying is primarily digital, geolocation metadata provides valuable contextual information. Location-based patterns can reveal concentrations of harmful behavior around specific schools, communities, or regional hotspots, enabling targeted intervention by local authorities. Correlating online aggression with offline events helps to understand real-world escalation dynamics, and in severe cases, geolocation data may contribute to physical safety assessments when digital threats imply immediate personal risk [38].

5.2. Integrated Use of Multimodal Data

A foundational principle of the platform architecture is that no single modality is sufficient for reliable cyberbullying detection. Many attacks are subtle and intentionally spread across different formats—text, visuals, audio, and contextual signals—to avoid detection. By integrating multimodal inputs, the system can differentiate isolated incidents from coordinated harassment, validate weak or ambiguous signals across data types, and track the temporal evolution of targeting behaviors. Multimodal machine learning models offer measurable improvements in precision and recall compared to unimodal approaches, while providing context-rich explanations for human moderators [33, 8]. This enables proactive defense, such as early detection of hostile hashtags, recognition of emerging meme clusters, and monitoring of repeated negative references to specific users.

Digital platforms can be conceptualized as social sensors: socio-technical infrastructures that continuously generate signals about human behavior, collective emotions, and emerging tensions. Unlike physical sensors that measure environmental variables such as temperature or movement, social sensors capture user-generated content and interactions including posts, comments, images, voice messages, reactions, and group dynamics. Within a Cyber Social Security (CSS) framework, these platforms operate as large-scale antennas capable of detecting early indicators of cyberbullying, hate speech, coordinated harassment, misinformation, and other cyber-social harms. Appropriately designed AI systems can transform this noisy and heterogeneous data stream into actionable insights for prevention, mitigation, and evidence-based policy support.

5.3. Data Types Collected by Social Sensors

The proposed SaaS platform adopts an explicitly multimodal approach, exploiting the diverse types of data emitted by online platforms. Each modality demands specialized computational techniques to effectively identify signals associated with cyberbullying and related phenomena.

5.3.1. Textual Data

Text remains the dominant interaction format across most platforms, including posts, tweets, comments, replies, private messages, and forum discussions. Through Natural Language Processing (NLP), combined with sentiment and toxicity analysis, the system can detect insults, slurs, threats, and humiliation, as well as coordinated attacks such as dogpiling. Analysis of conversational structure further enables identification of recurring patterns targeting particular individuals or groups. Additionally, monitoring emerging hashtags, recurrent expressions, and linguistic cues supports early-warning mechanisms capable of flagging the onset of targeted campaigns.

5.3.2. Visual Data: Images and Video

Platforms like Instagram, TikTok, Snapchat, and Facebook rely heavily on visual content, which has become a primary medium for cyberbullying and online hate. Relevant abusive behaviors include non-consensual sharing of photos, appearance-based mockery and body shaming, meme-based ridicule, and the disclosure of private imagery intended to shame individuals. Computer Vision methods enable recognition of harmful visual content by detecting explicit or suggestive imagery used for harassment,

reading embedded abusive text via Optical Character Recognition (OCR), and characterizing malicious memes or reaction videos directed at vulnerable groups. This capability is essential given that much contemporary aggression is encoded visually rather than verbally.

5.3.3. Audio and Voice Messages

Voice-based communication is increasingly prevalent in platforms such as Discord, Telegram, WhatsApp, Clubhouse, and online gaming environments. Through Automatic Speech Recognition (ASR) and audio analysis, the system can transcribe voice notes, detect verbal aggression including threats and humiliation, and evaluate prosodic features such as tone and intensity as additional indicators of hostility. This makes it possible to monitor abusive interactions in spaces that lack persistent text traces but nonetheless inflict substantial emotional harm.

5.3.4. Geolocated Data

While cyberbullying is primarily digital, geolocation metadata provides valuable contextual information. Location-based patterns can reveal concentrations of harmful behavior around specific schools, communities, or regional hotspots, enabling targeted intervention by local authorities. Correlating online aggression with offline events helps to understand real-world escalation dynamics, and in severe cases, geolocation data may contribute to physical safety assessments when digital threats imply immediate personal risk.

5.4. Integrated Use of Multimodal Data

A foundational principle of the platform architecture is that no single modality is sufficient for reliable cyberbullying detection. Many attacks are subtle and intentionally spread across different formats—text, visuals, audio, and contextual signals—to avoid detection. By integrating multimodal inputs, the system can differentiate isolated incidents from coordinated harassment, validate weak or ambiguous signals across data types, and track the temporal evolution of targeting behaviors. Multimodal machine learning models offer measurable improvements in precision and recall compared to unimodal approaches, while providing context-rich explanations for human moderators. This enables proactive defense, such as early detection of hostile hashtags, recognition of emerging meme clusters, and monitoring of repeated negative references to specific users.

In this framework, multimodality does not function solely as a technical enhancement but as the operational basis for a SAI model. Automated analysis provides continuous monitoring, large-scale pattern recognition, and rapid screening of vast data streams, while human moderators interpret contextual nuances, assess intent, apply ethical considerations, and make final escalation or intervention decisions. By distributing responsibilities according to the complementary strengths of machines and humans, the system achieves both scalability and responsible oversight, ensuring that detection remains accurate, transparent, and aligned with social and legal expectations.

5.5. Social Platforms as Concrete Sensors

Each social platform functions as a sensor with unique affordances, challenges, and risk levels. The proposed SaaS solution interfaces with multiple sources via APIs, scraping pipelines, or streaming connectors, adapting to the operational characteristics of each environment.

5.5.1. X (formerly Twitter)

X/Twitter represents a high-velocity, predominantly public microblogging environment where short textual posts can reach virality within minutes. Hashtags act as organizing tools for large-scale discussions, while public exposure intensifies the psychological effects of shaming and pile-ons. As a sensor, Twitter is well suited for real-time monitoring of hostile discourse, identification of harassment

campaigns targeting individuals or communities, and analysis of how influential users shape and amplify toxic narratives. The platform provides continuous streams of text and media suitable for early-warning systems.

5.5.2. Facebook and Instagram

Facebook and Instagram, both operated by Meta, combine strong relational structures with a mix of public, private, and semi-private communication channels including groups, pages, direct messages, and stories. Facebook groups may host organized harassment, while direct messaging enables abuse that is less visible. Instagram's strong focus on appearance and self-presentation makes it particularly prone to body shaming and image-based aggression. From a technical perspective, these platforms emphasize modeling group dynamics and considering ephemeral formats and recommendation algorithms that may inadvertently amplify toxic behavior.

5.5.3. Reddit

Reddit is structured around thematic communities (subreddits) characterized by pseudonymity and open debate. Such environments can foster aggressive expression due to lowered social inhibition, yet many communities also maintain robust self-moderation through rules and volunteer moderators. Reddit is particularly valuable for studying community-level toxicity emergence, detecting localized harassment patterns, and evaluating the interplay between moderation policies and harmful behaviors.

5.5.4. Telegram

Telegram supports large public groups, broadcast channels, and end-to-end encrypted secret chats, functioning both as a public sensor and a highly privacy-sensitive environment. Accordingly, automated monitoring must be restricted to legally and ethically accessible data sources, focusing on public or opt-in contexts. Telegram also illustrates adversarial adaptation, as malicious users may migrate or recreate groups following enforcement actions. This highlights practical trade-offs between privacy protection and cyberbullying prevention.

5.5.5. Other Platforms: TikTok, Snapchat, WhatsApp

TikTok emphasizes short-form participatory videos that can facilitate viral shaming, mocking edits, and discriminatory trends. Snapchat's ephemeral communication complicates retrospective monitoring but is relevant for real-time detection. WhatsApp and similar encrypted messengers are less suitable for broad sensing due to private group boundaries but are essential for localized, opt-in interventions such as school-based monitoring. Together, these platforms underscore the necessity of a unified multimodal SaaS solution capable of adapting to diverse ecosystem characteristics while sharing core analytical components including NLP, computer vision, ASR, and multimodal fusion techniques.

6. Multimodal AI Architecture for the Proposed SaaS Platform

Building on insights from the state of the art, we propose a cloud-native SaaS platform for multimodal detection and mitigation of cyber-social harms. The system operates across diverse digital environments, ingesting heterogeneous content streams and producing risk assessments, contextual explanations, and recommended actions for human operators.

6.1. Input Layer and Preprocessing

The platform ingests multimodal data including textual content from posts, comments, replies and direct messages (where lawfully accessible), as well as OCR-extracted text from images and video frames. Visual inputs comprise photographs, screenshots, memes and story frames, while video ingestion

includes short clips, live stream excerpts and gameplay fragments. Audio content such as voice notes and in-game communication is processed through segmentation and ASR transcription. Metadata including timestamps, platform identity, language, and basic relational context is also collected, with geolocation and network-structure features incorporated only where allowed by legal and ethical constraints. All signals are associated with a coherent event structure to support unified cross-modal analysis. Such comprehensive pre-processing pipelines are in line with recent guidelines for multimodal social media analysis [39].

6.2. Modality-Specific Encoders

Each modality is processed by a dedicated encoder designed to optimize accuracy, interpretability and inference cost.

6.2.1. Text Encoder

A Transformer-based encoder (e.g., BERT or derivative) fine-tuned on datasets related to cyberbullying, hate speech and coded abusive language generates both pooled sentence representations for classification and token-level embeddings supporting span-level explanations. This leverages the strong performance improvements of deep models over traditional ML approaches [40].

6.2.2. Visual Encoders: CLIP and Vision–Language Sequence Model

For visual content, the platform combines CLIP-style contrastive encoders [41] for scalable zero- or few-shot classification with a vision–language sequence-to-sequence model that supports region-level reasoning and contextual description. This duality enables both rapid triage and deeper forensic analysis of harmful visual behavior, addressing limitations identified in multimodal meme detection literature [15, 17].

6.2.3. Audio and Speech Encoder

The audio pipeline transcribes speech (via ASR) for textual analysis while extracting paralinguistic cues such as prosody and emotional tone, which provide additional evidence of hostility in voice-based exchanges. This modality is increasingly recognized as important for capturing harmful content beyond static image/text [17].

6.2.4. Metadata and Network Encoder

Where available, structured metadata and relational signals — e.g., user interaction history, group membership, follower/followee graphs — are encoded through numerical features and graph embeddings generated by Graph Neural Networks (GNNs) to capture social dynamics relevant to coordinated harassment or repeated abuse [42, 43].

6.3. Fusion and Decision Layer

The platform employs a flexible fusion strategy. For high-throughput screening tasks, a *late-fusion* scheme combines modality-specific predictions with explicit attribution of contributions. For detailed investigations, *intermediate fusion* with cross-attention captures subtle interdependencies between modalities, enabling detection of cases where single modalities appear benign but their combination yields harmful content. Ensemble-based mechanisms increase robustness against adversarial transformations, coded-language variation, and missing modalities. Similar architectures have been shown to outperform unimodal and simple fusion baselines on multimodal hate detection tasks [11, 15].

Outputs include continuous risk scores for multiple harm categories, severity levels, and recommended responses tailored to institutional practices.

6.4. Explainability and Human-in-the-Loop

Explainability mechanisms accompany each automated decision through text span highlighting, region-level visualization in images, key-frame selection in video, and contextual factors such as interaction history. Human moderators retain final authority to confirm, correct or escalate decisions. Feedback is fed into continual-learning pipelines, enabling adaptation to evolving language, new memes, and emerging harassment patterns. This hybrid human–AI design reflects recent calls for explainable, culturally aware and socially sensitive cyberbullying detection systems. [44, 43]

Such a SAI approach allows automated pattern recognition to scale, while preserving normative judgment and ethical oversight where context, culture, and human dignity matter most.

7. Evaluation and Case Studies

To assess feasibility, we plan evaluations on platforms with contrasting communication characteristics. For example, Reddit emphasizes community-centric, pseudo-anonymous discussions; Telegram offers group-based, partially closed interactions with strong multimedia components. This contrast allows us to test the platform across distinct social and technical environments, measuring detection accuracy, false positive/negative rates, human-in-the-loop latency, moderation workload, and model adaptability over time.

7.1. Experimental Setup

The evaluation pipeline consists of collecting posts, comments, images, and — where available — audio and video through scrapers and platform APIs. Conversations are reconstructed into hierarchical structures such as threads and reply chains. Samples are annotated for presence and subtype of cyberbullying or hate speech, as well as severity levels.

We compare multiple model configurations: textual encoders (e.g., RoBERTa-based models and larger LLM-derived classifiers), visual encoders such as CLIP and Florence-2–style vision-language models, ASR-based pipelines for speech, and multimodal ensembles integrating text, image, OCR and metadata features. This mirrors evaluation strategies adopted in recent multimodal hate detection research [2, 45].

7.2. Metrics

Performance is measured using precision, recall and macro-averaged F1-scores, alongside AUC-ROC for continuous risk scoring. Error analyses examine false positives and false negatives by harm category, platform, and modality, aligning with best practices in fairness-aware and explainable moderation research [46, 47].

7.3. Results Overview

Preliminary results indicate that multimodal ensembles substantially outperform unimodal baselines — especially in cases where sarcasm or visual cues dominate. On Reddit and Telegram datasets, binary bullying detection reaches F1-scores around 0.84 using CLIP + RoBERTa + OCR + ASR configurations on consumer-grade GPUs. Visual encoders significantly improve detection of appearance-based harassment and meme-centric abuse where harmful messaging is embedded directly in images — consistent with findings from multimodal hate-speech detection literature [2, 48].

Larger LLM-based text classifiers provide limited benefits in low-resource languages while incurring higher latency and compute costs — making them less suited for production-grade, real-time deployment.

7.4. Error Analysis

Common error patterns include: misinterpreting irony or self-directed humor as harmful content; under-detecting coded or community-specific language; and failing to capture prior relational context

that shifts interpretation. Detection performance also degrades on dialectal or mixed-language content due to limited language model coverage. These outcomes confirm that multimodality is necessary but insufficient by itself — reinforcing the importance of human oversight and culturally adaptive datasets, annotation protocols, and moderation policies [49, 47].

8. Discussion: Symbiotic AI, Human-in-the-Loop Workflow, and Responsible Deployment

The proposed SaaS platform is deliberately designed as *Symbiotic AI (SAI)*, where automation enhances rather than replaces human expertise. This design choice is essential in the context of cyberbullying and online hate, where cultural, relational, and ethical nuances require informed human judgment rather than purely algorithmic decisions [46, 50]. Consequently, the system supports moderators, educators, institutions, and platform operators through continuous multimodal monitoring and context-aware decision assistance [49].

From an operational perspective, SAI distributes responsibilities according to complementary strengths: artificial intelligence provides scalable, real-time analysis across vast and heterogeneous data streams, while humans retain oversight of interpretation, normative assessment, and high-impact interventions. This symbiosis enables efficient detection without sacrificing accountability, contextual sensitivity, or procedural fairness.

8.1. Roles and Stakeholders

Multiple human stakeholders interact with the platform. Trust & Safety teams within social media organizations review high-risk alerts, configure policies, and manage escalations. Educators and school staff receive contextualized reports related to minors or school communities, aligned with legal and institutional protocols. NGOs and support services are provided with anonymized insights into trends affecting vulnerable groups. In well-defined cases, law-enforcement or child protection authorities may access forensic-grade evidence — but always under due process and subject to privacy and data-protection constraints [46].

The AI system contributes by delivering ongoing surveillance across integrated platforms, multimodal risk scoring for different harm categories, and explanatory artifacts such as highlighted text spans, visual evidence, and interaction histories. The platform also incorporates feedback and correction mechanisms, enabling continual refinement of the models — a strategy shown to improve robustness, reduce performance drift, and strengthen trust in deployed moderation systems [50, 47].

8.2. Incident Lifecycle

A typical incident lifecycle begins with ingestion and pre-screening. New online events — posts containing images, early replies, audio snippets, or other signals — are collected through APIs or streaming connectors. Multimodal encoders generate preliminary risk assessments. Automated triage routes cases: low-risk events contribute to trend monitoring; medium-risk items enter human-review queues; high-risk events (e.g., explicit abuse involving minors) are immediately escalated [2, 49].

During human review, moderators access an explainability dashboard that highlights harmful text segments, problematic image regions or video frames, conversational context, and relational information about repeated offenders or targets. Reviewers may confirm or override automated labels, re-label content when necessary, and decide on actions such as warnings, content removal, or escalation.

All interventions feed into a continual-learning loop: misclassified cases become training signals; new forms of harm (e.g., emerging memes or coded language) are incorporated through few-shot prompting or targeted fine-tuning; and risk thresholds are periodically recalibrated based on error analyses and stakeholder feedback [50, 47]. This lifecycle operationalizes SAI by combining scalable machine perception with human judgment and responsibility.

8.3. Ethical, Legal, and Societal Considerations

Because the platform operates in a high-stakes domain affecting vulnerable individuals, ethical alignment and regulatory compliance are foundational design requirements rather than secondary constraints. Privacy and data protection obligations (e.g., GDPR) necessitate strict data minimization, purpose limitation, and pseudonymization wherever possible. Passive scanning of private or encrypted communications must be avoided except under lawful, opt-in, or tightly regulated conditions [46].

Fairness considerations are equally critical. Detection systems may inadvertently disadvantage dialects, minority groups, or reclaimed community-specific language. To mitigate these risks, the platform incorporates diverse training datasets, periodic fairness audits, and stakeholder participation in annotation and guideline development [47, 49].

False positives and false negatives each entail distinct harms: unjust sanctions may erode trust and chill legitimate speech, whereas missed abuse may enable escalation and victimization. For this reason, automated actions remain conservative, with human confirmation required for severe sanctions. Transparent logging, explainable decisions, and accessible appeal mechanisms further support accountability and procedural justice, consistent with emerging human-in-the-loop governance principles [50, 46].

Finally, the platform is designed to operate within existing legal and institutional frameworks through jurisdiction-specific policy profiles and auditable reporting mechanisms when escalations to authorities are mandated. By embedding ethical, legal, and societal safeguards directly into the workflow, the system ensures that technological scalability remains aligned with human values and democratic oversight.

9. Conclusion and Future Directions

This work introduced a multimodal, human-centered SaaS platform for detecting and mitigating cyber-social harms across online ecosystems. Drawing on recent advances in contrastive vision–language modeling, sequence-to-sequence multimodal reasoning, efficient text encoding, speech analysis, and human-in-the-loop moderation, the system offers a practical yet ethically grounded path to Cyber Social Security.

While initial experiments on Reddit and Telegram show promising results, limitations remain — including data imbalance, subtle relational contexts, computational costs of large models, and the inherent tension between safety, privacy, and freedom of expression. Future directions include domain adaptation for specific communities, improved performance in under-represented languages, lightweight architectures for efficient deployment, enhanced temporal reasoning to address evolving harassment patterns, real-time interventions with user-facing workflows, and participatory governance involving educators, youth, NGOs and marginalized populations. These developments are essential for building trustworthy, resilient, and inclusive digital spaces.

Acknowledgments

This work was supported by the following project: National Recovery and Resilience Plan, Mission 4 “Education and Research” – Component 2 “From Research to Business” – Investment 1.3, funded by the European Union – NextGenerationEU – CUP: F53C22000740007 – through the project titled “Cyber Social Security,” acronym CSS.

Declaration on Generative AI

During the preparation of this work, the author(s) used GPT-4 for grammar and spelling assistance. All content was reviewed and edited by the authors, who take full responsibility for the final publication.

References

[1] M. Calvano, A. Curci, G. Desolda, A. Esposito, R. Lanzilotti, A. Piccinno, Building symbiotic artificial intelligence: Reviewing the ai act for a human-centred, principle-based framework, Minds and Machines 36 (2026). URL: https://link.springer.com/article/10.1007/s11023-025-09753-w. doi:10.1007/s11023-025-09753-w.

[2] R. Prabhu, V. Seethalakshmi, A comprehensive framework for multi-modal hate speech detection in social media using deep learning, Scientific Reports 15 (2025) 13020.

[3] L. Cheng, J. Li, Y. N. Silva, D. L. Hall, H. Liu, Xbully: Cyberbullying detection within a multi-modal context, in: The 12th ACM International Conference on Web Search and Data Mining (WSDM '19), 2019, p. –. doi:10.1145/3289600.3291037.

[4] A. G. Philipo, D. S. Sarwatt, J. Ding, M. Daneshmand, H. Ning, Cyberbullying detection: Exploring datasets, technologies, and approaches on social media platforms, arXiv preprint (2024).

[5] H. Mehta, K. Passi, Social media hate speech detection using explainable artificial intelligence (xai), Algorithms 15 (2022) 291–313.

[6] T. M. et al., Artificial intelligence to address cyberbullying, harassment and abuse: A review of the state of the art, JMIR Publications (2022).

[7] A. A. et al., Image cyberbullying detection and recognition using deep learning + classical classifiers, Computers Security (2024).

[8] P. Vijayaraghavan, H. Larochelle, D. Roy, Interpretable multi-modal hate speech detection, in: arXiv preprint, 2021.

[9] A. Nirmal, A. Bhattacharjee, P. Sheth, H. Liu, Towards interpretable hate speech detection using large language model-extracted rationales, arXiv preprint (2024).

[10] E. F. et al., Enhancing social media hate speech detection in low-resource languages via transformer-based and explainable ai, Social Network Analysis and Mining (2025).

[11] D. Kiela, H. Firooz, A. Mohan, V. Goswami, A. Singh, P. Ringshia, D. Testuggine, The hateful memes challenge: Detecting hate speech in multimodal memes, in: Advances in Neural Information Processing Systems (NeurIPS) 2020, 2020.

[12] A. Das, J. S. Wahi, S. Li, Detecting hate speech in multi-modal memes, arXiv preprint arXiv:2012.14891 (2020).

[13] Y. Chen, et al., Multimodal detection of hateful memes by applying a vision–language pretraining model, PLOS ONE (2022).

[14] Y. Zhou, et al., Multimodal learning for hateful memes detection, arXiv preprint (2021).

[15] F. Wu, et al., Multimodal hateful meme classification based on transfer learning and clip, Electronics 13 (2024) 2780.

[16] A. Chhabra, D. K. Vishwakarma, Mhs-stma: Multimodal hate speech detection via scalable transformer-based multilevel attention framework, arXiv preprint arXiv:2409.05136 (2024).

[17] G. A. Koushik, D. Kanojia, H. Treharne, Towards a robust framework for multimodal hate detection: A study on video vs. image-based content, arXiv preprint (2025).

[18] E. Hossain, O. Sharif, M. M. Hoque, Mute: A multimodal dataset for detecting hateful memes, in: Proceedings of the 2nd Asia-Pacific Chapter of ACL, 2022, pp. 32–39.

[19] P. K. Smith, J. Mahdavi, M. Carvalho, S. Fisher, S. Russell, N. Tippett, Cyberbullying: Its nature and impact in secondary school pupils, Journal of Child Psychology and Psychiatry 49 (2008) 376–385.

[20] J. W. Patchin, S. Hinduja, Bullying Today: Bullet Points and Best Practices, Routledge, 2015.

[21] R. Slonje, P. K. Smith, A. Frisén, The nature of cyberbullying, and strategies for prevention, Computers in Human Behavior 29 (2013) 26–32.

[22] d. boyd, M. Ito, et al., Hanging Out, Messing Around, and Geeking Out: Kids Living and Learning with New Media, MIT Press, 2010.

[23] R. M. Kowalski, S. P. Limber, P. W. Agatston, Bullying in the Digital Age, 2 ed., Wiley-Blackwell, 2014.

[24] Y. Huang, H. Chou, The role of online bystanders in cyberbullying incidents: Motivations and behavior, Computers & Education 178 (2022) 104399.

[25] R. M. Kowalski, S. P. Limber, Cyberbullying and adolescent mental health: Systematic review of the literature, Journal of Adolescent Health 53 (2012) S13–S20.
[26] M. P. e. a. Hamm, Prevalence and effect of cyberbullying on children and young people: A scoping review, JAMA Pediatrics 169 (2015) 770–777.
[27] S. Bauman, D. Cross, J. Walker, Principles of Cyberbullying Research, Routledge, 2013.
[28] P. Voigt, A. Von dem Bussche, The EU General Data Protection Regulation (GDPR): A Practical Guide, Springer, 2017.
[29] S. Livingstone, L. Haddon, A. Görzig, K. Ólafsson, Eu kids online: Final report (2011).
[30] Z. Waseem, T. Davidson, D. Warmsley, I. Weber, Understanding abuse: A typology of abusive language detection subtasks, in: Proceedings of the First Workshop on Abusive Language Online, 2017, pp. 78–84.
[31] T. Sakaki, M. Okazaki, Y. Matsuo, Earthquake shakes twitter users: Real-time event detection by social sensors, in: WWW, 2010.
[32] D. e. a. Lazer, Computational social science, Science 323 (2009) 721–723.
[33] T. Baltrušaitis, C. Ahuja, L.-P. Morency, Multimodal machine learning: A survey and taxonomy, IEEE TPAMI (2019).
[34] T. e. a. Davidson, Automated hate speech detection and the problem of offensive language, in: ICWSM, 2017.
[35] A. Schmidt, M. Wiegand, A survey on hate speech detection, Natural Language Engineering (2017).
[36] B. W. Schuller, Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends, Communications of the ACM 61 (2018) 90–99. doi:10.1145/3129340.
[37] A. Baevski, Y. Zhou, A. Mohamed, M. Auli, wav2vec 2.0: A framework for self-supervised learning of speech representations, in: Advances in Neural Information Processing Systems (NeurIPS), volume 33, 2020, pp. 12449–12460.
[38] A. Crooks, A. Croitoru, A. Stefanidis, J. Radzikowski, earthquake: Twitter as a distributed sensor system, Transactions in GIS 19 (2015) 124–147. doi:10.1111/tgis.12108.
[39] T. T. Nguyen, et al., Decoding digital discourse through multimodal text and media: Practices and challenges, Journal of Medical Internet Research (JMIR) (2025).
[40] M. T. Hasan, et al., A review on deep-learning-based cyberbullying detection, cognitive systems research (2023).
[41] A. El-Sayed, O. Nasr, Aast-nlp at multimodal hate speech event detection 2024: A multimodal approach for classification of text-embedded images based on clip and bert-based models, in: Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2024), Association for Computational Linguistics, 2024, pp. 139–144.
[42] K. Maity, T. Sen, S. Saha, P. Bhattacharyya, Mtbullygnn: A graph neural network–based multitask framework for cyberbullying detection, IEEE Transactions on Computational Social Systems (2022).
[43] G. M. Mohiuddin, et al., Deep learning models for culturally aware cyberbullying detection, Multimedia Tools and Applications (2025).
[44] S. García-Méndez, F. de Arriba-Pérez, Explainable cyberbullying detection using large language models in a stream-based ml framework, arXiv preprint (2025).
[45] M. Das, R. Raj, P. Saha, B. Mathew, M. Gupta, A. Mukherjee, Hatemm: A multi-modal dataset for hate video classification, arXiv preprint arXiv:2305.03915 (2023).
[46] X. Wang, et al., The unappreciated role of intent in algorithmic moderation of abusive content on social media, MISinfoReview (2025).
[47] E. Fetahi, et al., Enhancing social media hate speech detection in low-resource languages via transformer-based models and explainable ai, Social Network Analysis and Mining (2025).
[48] P. Lippe, N. Holla, S. Chandra, S. Rajamanickam, G. Antoniou, E. Shutova, H. Yannakoudakis, A multimodal framework for the detection of hateful memes, in: arXiv preprint arXiv:2012.12871, 2020.

[49] S. Udupa, et al., Recent advances in online hate speech moderation: Multimodality and the role of large models, Computational Social Science Review (2023).

[50] A. Kotarčić, D. Hangartner, F. Gilardi, S. Kurer, K. Donnay, Human-in-the-loop hate speech classification in a multilingual context, Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (2022).